



# Fully Automated and Accurate Annotation of Eukaryotic Genomes with BRAKER & TSEBRA

RESPONSE Summer School 2022

Katharina J. Hoff,  
Lars Gabriel,  
Tomáš Brůna,  
Alexandre Lomsadze,  
Mario Stanke,  
Mark Borodovsky

Contact: [katharina.hoff@uni-greifswald.de](mailto:katharina.hoff@uni-greifswald.de)

Katharina J. Hoff,  
Lars Gabriel,  
Tomáš Brůna,  
Alexandre Lomsadze,  
Mario Stanke,  
Mark Borodovsky



Gene Prediction

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

Summary

References

# Contents

## 1 Gene Prediction

## 2 BRAKER1: RNA-Seq

## 3 BRAKER2: Proteins

## 4 TSEBRA

## 5 Summary

## Structural Genome Annotation Problem

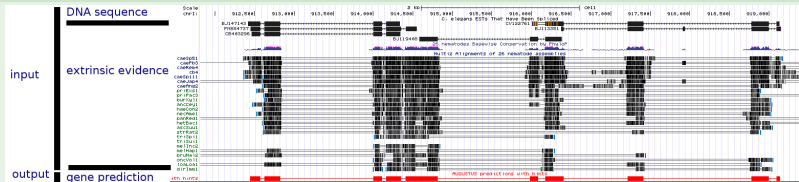
## Input

- **genome** assembly
- **extrinsic evidence**, e.g. from RNA-Seq, **protein database**

## Output

- protein-coding genes: exon-intron structures (.gff)

### Example (from Chr I in *C. elegans*)

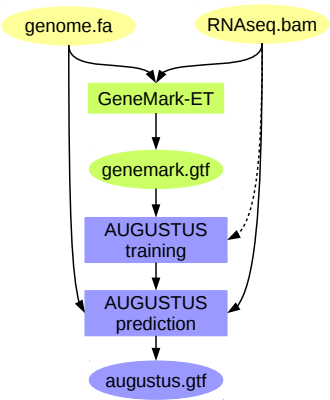




# BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS FREE

Katharina J. Hoff ✉, Simone Lange, Alexandre Lomsadze, Mark Borodovsky ✉, Mario Stanke

Bioinformatics, Volume 32, Issue 5, 1 March 2016, Pages 767–769,  
<https://doi.org/10.1093/bioinformatics/btv661>



- spliced alignments of RNA-Seq are used by GeneMark-ET and AUGUSTUS
- 721 citations (Google Scholar)

## Whole-Genome Annotation with BRAKER

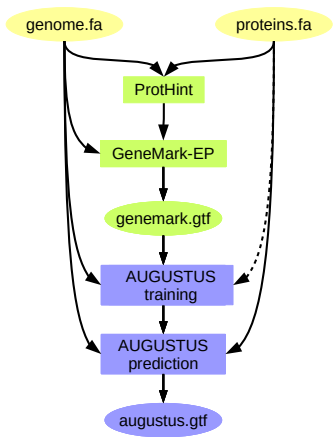
Katharina J. Hoff, Alexandre Lomsadze, Mark Borodovsky,  
and Mario Stanke

in Kollmar M. (eds) Gene Prediction. Methods in Molecular Biology,  
vol 1962. Humana, New York, NY, 2019



# BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database

Tomáš Brůna<sup>1,†</sup>, Katharina J. Hoff<sup>2,3,†</sup>, Alexandre Lomsadze<sup>4</sup>, Mario Stanke<sup>2,3,‡</sup> and  
Mark Borodovsky<sup>4,5,\*</sup>



- spliced alignments of a large number of proteins are used by GeneMark-EP and AUGUSTUS
- 220 citations (Google Scholar)

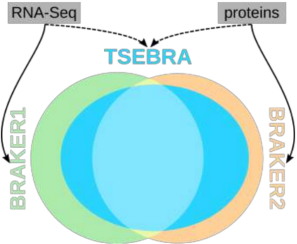


SOFTWARE Open Access



# TSEBRA: transcript selector for BRAKER

Lars Gabriel<sup>1,2</sup>, Katharina J. Hoff<sup>1,2</sup>, Tomáš Brůna<sup>3</sup>, Mark Borodovsky<sup>4,5</sup> and Mario Stanke<sup>1,2\*</sup>



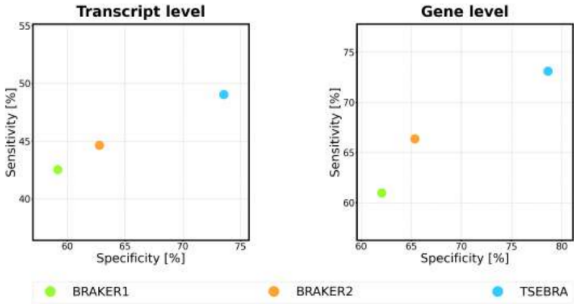
- BRAKER annotation with RNA-Seq and proteins
- achieve higher accuracy than BRAKER1 and BRAKER2
- 10 citations (Google Scholar)

Katharina J. Hoff,  
Lars Gabriel,  
Tomáš Brůna,  
Alexandre Lomsadze,  
Mario Stanke,  
Mark Borodovsky



# Average prediction accuracy

- *Arabidopsis thaliana*
- *Caenorhabditis elegans*
- *Drosophila melanogaster*



Katharina J. Hoff,  
Lars Gabriel,  
Tomáš Brůna,  
Alexandre Lomsadze,  
Mario Stanke,  
Mark Borodovsky



Gene Prediction

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

Summary

References

## Summary

- fully automatic pipelines
- state-of-the art accuracy
- fast & easy to use

## Ongoing Development

- BRAKER3 with built-in RNA-Seq & protein support
- Machine learning based TSEBRA





## References

- Gabriel L et al. (2021) “TSEBRA: transcript selector for BRAKER”
- Bruna T et al. (2021) “BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database”
- Hoff KJ et al. (2016) “BRAKER1: unsupervised RNAseq-based genome annotation with GeneMark-ET and AUGUSTUS.”
- Hoff KJ et al. (2019) “Whole-genome annotation with BRAKER.”
- Lomsadze A et al. (2014) “Integration of mapped RNAseq reads into automatic training of eukaryotic gene finding algorithm.”
- Stanke M et al. (2008) “Using native and syntenically mapped cDNA alignments to improve de novo gene finding.”
- Bruna T et al. (2020) “GeneMark-EP and -EP+: automatic eukaryotic gene prediction supported by spliced aligned proteins.”
- Buchfink B et al. (2015) “Fast and sensitive protein alignment using DIAMOND.”
- Iwata H & Gotoh O (2012) “Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features.”

**BRAKER & TSEBRA are Available for Download at**

- <https://github.com/Gaius-Augustus>

Katharina J. Hoff,  
Lars Gabriel,  
Tomáš Brůna,  
Alexandre Lomsadze,  
Mario Stanke,  
Mark Borodovsky



Gene Prediction

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

Summary

References

## Co-Authors

Lars Gabriel  
Simone Lange  
Alexandre Lomsadze  
Tomáš Brůna  
Mario Stanke  
Mark Borodovsky

## Funding

This research is supported by  
US National Institutes of Health  
grant GM128145 to Mark  
Borodovsky and Mario Stanke.

## Acknowledgements

Hannah Thierfeldt  
Anica Hoppe  
Jens Keilwagen  
Maria Hartmann  
Ingo Bulla  
Timon Kapischke  
Holger Irrgang  
Felix Becker  
Matthis Ebel

**BRAKER & TSEBRA are Available for Download at**

- <https://github.com/Gaius-Augustus>

Katharina J. Hoff,  
Lars Gabriel,  
Tomáš Brůna,  
Alexandre Lomsadze,  
Mario Stanke,  
Mark Borodovsky



Thank you for your attention!

Gene Prediction

BRAKER1: RNA-Seq

BRAKER2: Proteins

TSEBRA

Summary

References

## **BRAKER2 is Available for Download at**

- <https://github.com/Gaius-Augustus/BRAKER>
- <https://github.com/gatech-genemark/BRAKER2>